Extensive Genetic Diversity and Substructuring Among Zebrafish Strains Revealed through Copy Number Variant Analysis

Extensive genetic diversity and substructuring among zebrafish strains revealed through copy number variant analysis

Kim H. Brown[a,b,1], Kimberly P. Dobrinski[a,b,1], Arthur S. Lee[a,2], Omer Gokcumen[a,b], Ryan E. Mills[a,b], Xinghua Shi[a,b], Wilson W. S. Chong[a,c], Jin Yun Helen Chen[a], Paulo Yoo[a], Sthuthi David[a], Samuel M. Peterson[d], Towfique Raj[b,e,f,g], Kwong Wai Choy[c], Barbara E. Stranger[b,e,f], Robin E. Williamson[h], Leonard I. Zon[i], Jennifer L. Freeman[a,b,3,4], and Charles Lee[a,b,f,4,5]

[a]Department of Pathology, Brigham and Women's Hospital, Boston, MA 02115; [b]Harvard Medical School, Boston, MA 02115; [c]Department of Obstetrics and Gynaecology, Chinese University of Hong Kong, Hong Kong SAR, China; [d]School of Health Sciences, Purdue University, West Lafayette, IN 47907; [e]Division of Genetics, Brigham and Women's Hospital, Boston, MA 02115; [f]Program in Medical and Population Genetics, Broad Institute, Cambridge, MA 02142; [g]Department of Neurology, Brigham and Women's Hospital, Boston, MA 02115; [h]Department of Obstetrics, Gynecology and Reproductive Biology, Harvard Medical School, Boston, MA 02115; and [i]Stem Cell Program and Division of Hematology/Oncology, Children's Hospital and Dana-Farber Cancer Institute, Howard Hughes Medical Institute, Harvard Stem Cell Institute, Harvard Medical School, Boston, MA 02115

Edited by Igor B. Dawid, National Institute of Child Health and Human Development, National Institutes of Health, Bethesda, MD, and approved November 22, 2011 (received for review July 26, 2011)

Copy number variants (CNVs) represent a substantial source of genomic variation in vertebrates and have been associated with numerous human diseases. Despite this, the extent of CNVs in the zebrafish, an important model for human disease, remains unknown. Using 80 zebrafish genomes, representing three commonly used laboratory strains and one native population, we constructed a genome-wide, high-resolution CNV map for the zebrafish comprising 6,080 CNV elements and encompassing 14.6% of the zebrafish reference genome. This amount of copy number variation is four times that previously observed in other vertebrates, including humans. Moreover, 69% of the CNV elements exhibited strain specificity, with the highest number observed for Tubingen. This variation likely arose, in part, from Tubingen's large founding size and composite population origin. Additional population genetic studies also provided important insight into the origins and substructure of these commonly used laboratory strains. This extensive variation among and within zebrafish strains may have functional effects that impact phenotype and, if not properly addressed, such extensive levels of germ-line variation and population substructure in this commonly used model organism can potentially confound studies intended for translation to human diseases.

comparative genomic hybridization | structural variation | gene expression

The zebrafish, *Danio rerio*, is an important model system for human pathologies, given its distinct advantages including high fecundity, external fertilization, and the availability of forward (i.e., induced random mutation) and reverse (i.e., gene knockout) genetic techniques (1). Although the zebrafish reference genome is now considered complete, it is still poorly characterized with respect to genetic variants [i.e., single-nucleotide polymorphisms (2) and structural genomic variants]. The limited number of population-wide analyses available (i.e., microsatellite polymorphisms) indicate that native zebrafish populations have significantly more variability than commonly used laboratory strains (3). This finding suggests that laboratory strains of zebrafish, compared with native zebrafish populations, should have considerably lower levels of structural variants.

Structural genomic variants, including both balanced (e.g., most inversions, insertions, and translocations) and unbalanced rearrangements (i.e., copy number variants; CNVs), are widespread among vertebrates, with CNVs representing the largest known component (4–11). In addition, CNVs have also been associated with human disease phenotypes including neurological disorders (12), early onset obesity (13), and some forms of cancer (14). Although a few zebrafish CNVs have been studied previously [e.g., mannose binding lectin (15) and globin (16) genes], no comprehensive genome-wide assessment for zebrafish CNVs has yet been reported. We have conducted an assessment revealing an unprecedented high level of copy number polymorphisms distributed throughout the genome, with extensive within- (intra-) and between- (inter-) strain variation. This variation indicates extensive genetic substructuring between zebrafish strains, with 69% of CNV elements (CNVEs; e.g., CNVs having >50% reciprocal overlap) exhibiting strain specificity, with the highest levels observed for Tubingen (Tu). The CNVE substructuring observed among laboratory strains appears to be primarily driven by the initial population size and the genetic variation within the founding stock. This variation has the potential to impact natural and experimentally derived phenotypes within the species, and could confound zebrafish studies intended for translation to human disease if not addressed by study designs.

Results and Discussion

Extensive Genomic Copy Number Variation Within and Between Laboratory Strains and a Native Zebrafish Population. We examined 80 individual fish from three separate zebrafish laboratory strains and a native population to identify and characterize CNV content within and between zebrafish lineages using array-based comparative genomic hybridization (aCGH) (Fig. 1). Twenty randomly selected fish were collected for each of the three laboratory strains AB, WIK, and Tu, as well as a native population from Bangladesh. Among laboratory strains, AB and Tu are commonly used for mutagenesis and gene knockdown experiments. WIK is predominantly used to facilitate gene mapping, because of its high level of simple sequence length polymorphism

Author contributions: K.H.B., K.P.D., J.L.F., and C.L. designed research; K.H.B., K.P.D., A.S.L., W.W.S.C., J.Y.H.C., P.Y., S.D., S.M.P., and J.L.F. performed research; T.R., K.W.C., and B.E.S. contributed new reagents/analytic tools; K.H.B., K.P.D., O.G., R.E.M., X.S., R.E.W., and L.I.Z. analyzed data; and K.H.B., K.P.D., and C.L. wrote the paper.

The authors declare no conflict of interest.

This article is a PNAS Direct Submission.

Data deposition: The data reported in this paper have been deposited in the Gene Expression Omnibus (GEO) database, www.ncbi.nlm.nih.gov/geo (reference super-series GSE28328).

[1]K.H.B. and K.P.D. contributed equally to this work.

[2]Present address: Washington University School of Medicine, St. Louis, MO 63110.

[3]Present address: School of Health Sciences, Purdue University, West Lafayette, IN 47907.

[4]J.L.F. and C.L. contributed equally as senior authors.

[5]To whom correspondence should be addressed. E-mail: clee@rics.bwh.harvard.edu.

This article contains supporting information online at www.pnas.org/lookup/suppl/doi:10.1073/pnas.1112163109/-/DCSupplemental.

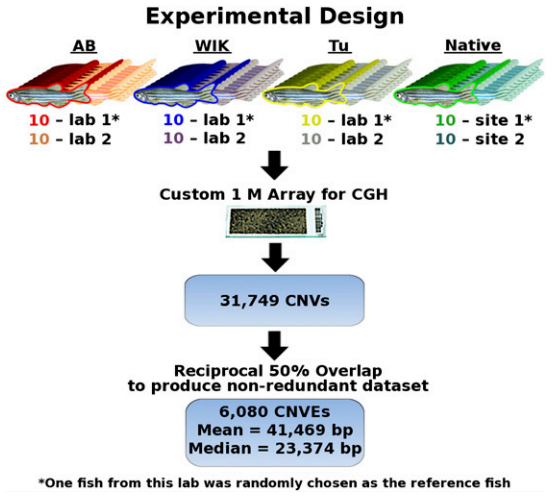

Fig. 1. Experimental design overview. Twenty zebrafish individuals from each of AB, WIK, and Tu and a native strain from Bangladesh were analyzed (10 individuals per laboratory site). A single individual from each strain was randomly selected as the strain reference to compare against remaining individuals in the strain using a custom Agilent 1 Million feature aCGH array platform. Analyses produced 31,749 CNVs that were combined into 6,080 CNVEs using a 50% reciprocal overlap criterion.

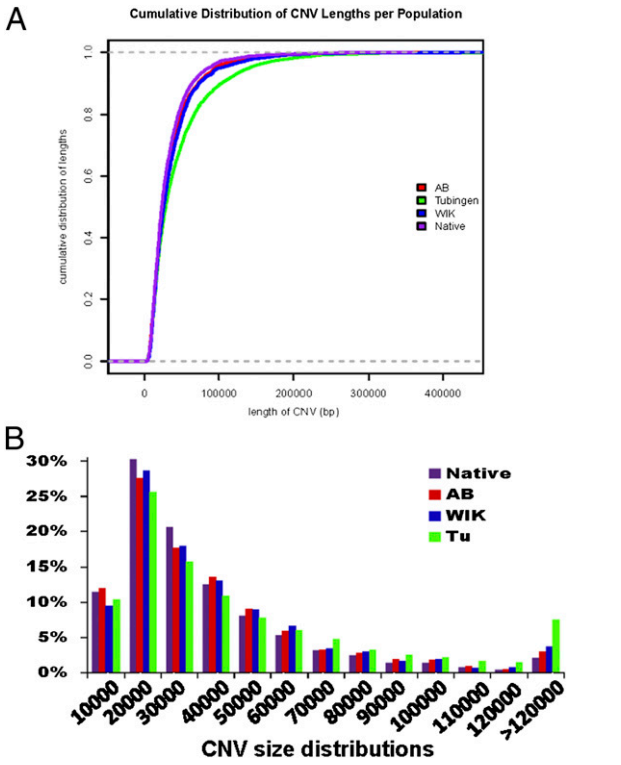

Fig. 2. CNV size variation. (*A*) Cumulative distribution function plots of CNV sizes for each strain. Faster-rising lines indicate an increased frequency of smaller CNVs, and slower-rising lines indicate an increased frequency of larger CNVs. Differences between strains were significant, with Tu having significantly larger CNVs and the native fish having significantly smaller CNVs (ANOVA; $P < 0.001$). (*B*) CNV frequency histogram of the percentage of total CNV calls within 10-kb size bins indicating a significantly higher percentage of small CNVs in native fish and a significantly higher percentage of larger CNVs in Tu (ANOVA; $P < 0.001$).

(e.g., microsatellites) variation compared with either AB or Tu (17). The aCGH experiments were performed using a custom array containing 967,331 uniquely mapping 60-mer oligonucleotide array features (i.e., test probes) designed against the zebrafish Zv8 reference genome, originating from the Tu strain. Unique probes were selected to reduce noise generated from highly repetitive regions, and as such our array lacks information for known zebrafish genome duplications (e.g., globin genes) and genomic regions containing segmental duplications, long interspersed nuclear elements (LINEs), short interspersed nuclear elements (SINEs), and simple repeats. Although we primarily avoided such sequences to reduce noise in our arrays, we also avoided these areas as many of the repeat elements (i.e., segmental duplications) have not yet been accurately mapped in the zebrafish genome. To minimize potential biases arising from the use of the Tu reference, we used strain-specific references for all groups.

CNV calls were based on mean $\log_2$ ratios of ± 0.4 for three consecutive probes. Using these criteria, we interrogated the zebrafish genome for relative DNA gains and losses at an effective resolution of ~4 kb. We attribute the observed $\log_2$ ratio differences between individuals to actual CNVs, because SNPs in zebrafish are thought to occur approximately once every 0.5–3 kb (2, 18), and at least three SNPs within a 60-mer are required to sufficiently alter probe binding efficiency on our array platform. These experiments identified a total of 31,749 CNVs across all 80 fish (Fig. 1). CNV size divergence between strains showed appreciable differences, with the native fish having significantly smaller CNVs (mean CNV size of 32,452 bp) compared with all other groups (ANOVA; $P < 0.001$), and Tu (mean CNV size of 45,731 bp) having significantly larger CNVs compared with all other groups (Tukey HSD; $P < 0.01$). The size distribution frequencies can be appreciated in a size frequency histogram and cumulative distribution function plot (Fig. 2). Among these CNVs, 100 were randomly selected for subsequent quantitative PCR (qPCR) validation experiments, with 95% of the loci tested subsequently validating (Table S1).

Pairwise analyses of genomic CNV coverage (Table S2) across the four zebrafish groups demonstrated a much higher level of variation (1.22%) than interindividual CNV differences found among human populations (0.78%) (4). This level of coverage only takes into account pairwise variation given the use of strain-specific reference DNAs. Combining all zebrafish CNVs discovered, a nonredundant dataset comprising 192,460,331 bp of sequence, representing 14.6% of the zebrafish reference genome, was obtained. This dataset represents more than four times the percentage of reference genome sequence covered by similarly common CNVs in humans (4) and other vertebrates (5, 8, 9). Reporting the percentage of the genome affected by CNVs should compensate for differences in genome sizes and array resolutions from different CNV studies. Although our array has a reduced resolution compared with some recent human studies (4), comparisons with similar resolution human arrays (10) still indicate that the content of CNVs in the zebrafish genome far exceeds that found in humans. Moreover, our array design precluded repetitive genomic elements (i.e., segmental duplications, LINEs, SINEs, etc.), which have been shown to be catalysts of CNVs in humans (19), suggesting that our CNV estimates are actually underrepresentations of the true amount of structural variation between strains.

To identify strain-specific CNV differences, additional aCGH experiments were conducted using pooled DNA of 10 additional fish each from AB, WIK, and Tu. These experiments identified 2,393 CNVs covering 58,930,737 bp (4.5%) of the zebrafish reference genome. Of these 2,393 CNVs, 682 (28.5%) did not overlap with CNVs discovered during the analysis of the original 80 fish. Of these CNVs, 162 were found to be strain-specific, with seven genes completely overlapped by CNVs (three Tu gains, zgc:163079, zgc:1629, *tmem103*; two Tu losses, *trh1*, *irx3a*; two WIK gains, zgc:77058, *ppp1r3b*) and 56 genes partially covered by CNVs (Table S3). Given that we used strain-specific refer-

ences for calling CNVs in each of the 80 individual fish, these observations indicate that our findings are underestimates of the total level of structural variation that actually exists between zebrafish strains.

The 31,749 CNVs found in the 80 individual fish were combined (using a 50% reciprocal overlap criterion) into a nonredundant dataset of 6,080 CNVEs (Dataset S1). The CNVEs exhibited a mean and median size of 41.5 kb and 23.4 kb, respectively, and appeared to be distributed ubiquitously across the 25 zebrafish chromosomes (Fig. 3). Two chromosomes, 4 and 22, were observed to have a greater percentage of their chromosome lengths associated with CNVEs (33.8% and 28.1%, respectively; chromosome average 14.2%; Fig. S1). These two chromosomes are known to have a high degree of heterochromatin and genic sequence, respectively. Additionally, we found 2,244 (37%) CNVEs overlapping 2,865 (19.4%) National Center for Biotechnology Information (NCBI) Reference Sequence (RefSeq) genes (Table 1). This number represents a significant depletion in the number of CNVEs expected to overlap RefSeq genes by chance alone ($P < 0.001$; Fig. S2*A* and *SI Text*). Also, similar to other vertebrate CNV studies (4, 20, 21), we found a significant enrichment for immunity-related genes (e.g., MHC class I UBA and chemokine ligand 12b) ($P < 0.001$; Fig. S2*B* and *SI Text*). Among overlapped RefSeq genes, 279 (9.7%) had CNVEs exclusively located within introns (Table 1). For the remaining 2,586 genes, CNVEs completely or partially overlapped the 5′ UTR or 3′ UTR, and eliminated exonic sequences potentially altering regulatory regions affecting gene expression (Table 1).

To determine whether CNVs in zebrafish influence gene function, we collected DNA and RNA from seven additional full-sib adult zebrafish. The impact of the 792 most copy number-variable CNVs then was assessed for correlative analyses on gene expression levels from RNA expression microarrays (*SI Text*). A *cis*-expression quantitative trait loci (eQTL) analysis was performed associating CNV copy numbers found within a 1-Mb window upstream and downstream of a gene's transcription start site and gene expression levels for that gene using a Spearman rank correlation (22). To assess the significance of nominal *P* values, we used the *q*-value false discovery rate (FDR) estimation, setting the FDR at 0.05 (23). In total, $\log_2$ ratios of 15,137 CNV probes were tested against expression levels of 11,953 gene probes, with 301 significant CNV probe–gene probe expression associations identified (Table S4) comprising 232 CNVs (29.3% of those tested) and 255 genes (2.5% of known genes). Because of the limited sample size for this analysis, we focused on broad patterns of detected associations, as opposed to specific individual associations. Associations were classified based on four parameters: direct associations (i.e., CNV overlapping RefSeq genomic sequence), indirect associations (i.e., CNV not overlapping RefSeq genomic sequence), positive associations (i.e., copy gain associated with increased expression or copy loss associated with decreased expression), and negative associations (i.e., a gain associated with decreased expression or a loss associated with increased expression). Seventy six (25.2%) of the associations were direct with positive associations, 40 (13.3%) were direct with negative associations, 104 (34.6%) were indirect with positive associations, and 81 (26.9%) were indirect with negative associations. Interestingly, the majority of CNV–gene expression associations were indirect (185; 61.5%) and therefore possibly regulatory in nature (24). These patterns reflect associations for only high copy-number variable CNVs, which may differ from that of the entire CNV set. We also note that these associations may not actually be attributable to the CNV itself, as a CNV could also serve as a proxy for a functional SNP. Nevertheless, this analysis suggests that many CNVs are likely to contribute to gene expression variation among zebrafish individuals and possibly to higher-order phenotypes, and motivates a more comprehensive, well-powered eQTL study to characterize specific patterns of functional effects of CNVs on gene expression in zebrafish.

Highest Genetic Variation Found Within the Tu Strain. Zebrafish are indigenous throughout the southeastern Himalayan region from Pakistan to Myanmar, at elevations near sea level to more than 1,300 m (25). Populations are typically found in slow-moving water with temperatures ranging between 20 and 34 °C (25). Although no extensive native population genetic studies have yet been performed, their widespread geographical range and diverse environmental conditions suggest local adaptation may occur between isolated groups. This widespread distribution could lead to substantial genetic substructuring among zebrafish populations, as has been observed previously for other freshwater aquatic species with disjunct distributions (26, 27). Indeed, our CNVE data appear to be consistent with extensive population substructuring (i.e., local adaptation) among zebrafish populations, with 4,199 (69%) of the identified CNVEs unique to one strain and only 457 (7.5%) CNVEs common to all four groups (Fig. 4*A*). To determine the significance of the apparent substructuring, we analyzed the CNVEs using FRAPPE (28). This analysis resulted in $K = 4$ as the most probable structure for the zebrafish strains (Fig. 4*B*). Although the genetic substructuring in zebrafish is not unexpected, it does indicate that care must be taken when examining zebrafish data, as some CNVs may cause strain-specific phenotypes.

The variation among the four zebrafish groups studied was most pronounced in Tu, which exhibited the highest level of CNV sequence coverage and genic CNVE coverage (Table 1). This finding somewhat contradicts previous microsatellite-based studies (3) suggesting that genetic variation within laboratory strains is considerably lower than among individuals of native zebrafish populations. Because our array platform was based on a Tu reference sequence, one limitation of our platform could be diminished sensitivity for detecting certain CNVs in other

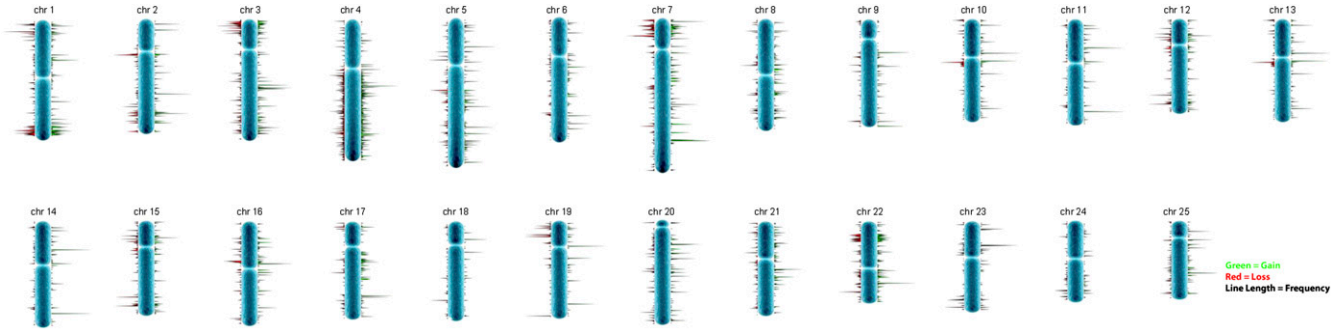

Fig. 3. CNV map. A combined zebrafish CNVE map with copy number gains (green) and losses (red) distributed along chromosome lengths. Length of green and red lines reflect relative CNVE frequencies at respective chromosomal locations.

Table 1. RefSeq genes affected by CNVs

Strain	Genes	Whole gene	Intronic	5′ UTR	3′ UTR	Other*
AB	1,094	677	75	133	123	86
WIK	1,091	682	85	150	133	41
Native	955	629	92	107	82	45
Tu	2,151	1,401	95	290	254	111
Total	2,865	1,802	279	409	351	24

Overlap analyses were performed to determine the number of RefSeq genes and gene components that overlap CNVs within each zebrafish strain. *Denotes genes with CNVs in exons or multiple CNVs affecting the 5′ UTR and 3′ UTR separately.

zebrafish strains. For example, DNA sequences that are only present in Tu would not be detected as homozygous losses when the test and reference DNAs are from the same non-Tu zebrafish strain. For Tu-specific sequences, non-Tu samples would result in no dye intensity signal for both the test and reference, theoretically resulting in a $\log_2$ ratio of zero. Alternatively, sequences completely absent from Tu but present in one or more copies in AB or WIK would not be represented on our array platform, and therefore CNVs for these genomic regions in AB and WIK would go undetected. Based on the number of highly confident homozygous losses that we detect in the zebrafish strains studied, we believe that the loss of CNV detection resulting from these limitations would be minimal.

Observed genotypic differences among laboratory strains likely result from at least two separate factors: initial population size and initial founding stock genetic variation. WIK (originally WIK11) was created from a single pair of native caught fish from India (17). The current AB strain (also known as AB*) was reinitiated in 1992 using six pairs. (AB strain history is available at http://zfin.org/action/genotype/detail?genotype.zdbID=ZDB-GENO-960809-7.) In contrast to these two strains, Tu originated from about 100 commercially raised founders (29) obtained from multiple locally adapted populations (i.e., composite populations). In such a composite population, initial fish are obtained from distinct, geographical locations and exhibit increased levels of genetic and presumably phenotypic variation (Fig. 5). Random mating between such founders results in hybrid offspring with greater levels of genetic variation than either parent. Continued random mating among hybrid fish, along with the increase in genetic recombination known to occur in composite populations (reviewed in ref. 30), would substantially increase genetic variation in subsequent offspring. Given Tu's origins from a composite population and its increased mean CNV size, it is suspected that Tu may harbor more segmental duplications than other laboratory strains as a result of hybridizations between founders with unique chromosomal rearrangements and/or duplication events. Specific chromosomal rearrangements have previously been observed between different zebrafish strains (e.g., AB and Tu) (31), and therefore could have existed among some of the Tu founders that were locally adapted. An increased number of segmental duplications in the Tu composite population would suggest an increased rate of nonallelic homologous recombination and hence the observed increases in CNV number and size among these fish compared with other zebrafish laboratory strains with more restricted lineages (Fig. 2).

To explore potential strain-specific characteristics, we both examined gene enrichments among CNVs and compared V_{st} values, a population differentiation estimator similar to F_{st} (10), among strains. For gene enrichment analyses, we conducted pairwise permutations between the four zebrafish groups for all CNV genic regions. These analyses identified 71 genes (Tu, 39; Native, 19; WIK, 9; AB, 4) exhibiting significantly enriched coverage within a strain (Fisher exact test; $P > 0.05$) with 55% observed only in Tu (Table S5). Calculations for V_{st} across all zebrafish groups were performed using aCGH $\log_2$ ratio signal intensity data. V_{st} data provide values from 0 (no difference) to 1

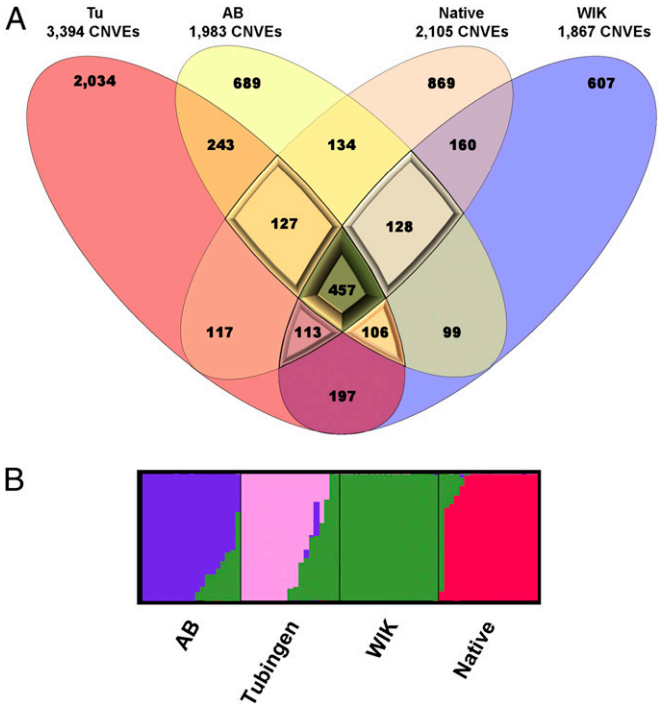

Fig. 4. Strain-specific differences. (*A*) A Venn diagram indicating strain-specific CNVEs and the numbers of overlapping CNVEs between strains. CNVEs observed in three or more strains are represented as raised shallow plateaus and three-way overlaps, with the highest plateau representing CNVEs occurring in all strains. (*B*) Structure plot of CNVE data analyzed by FRAPPE for population substructure found an optimal value of $K = 4$. Analyses were also performed for $K = 5$–8 with no significant changes in the structure.

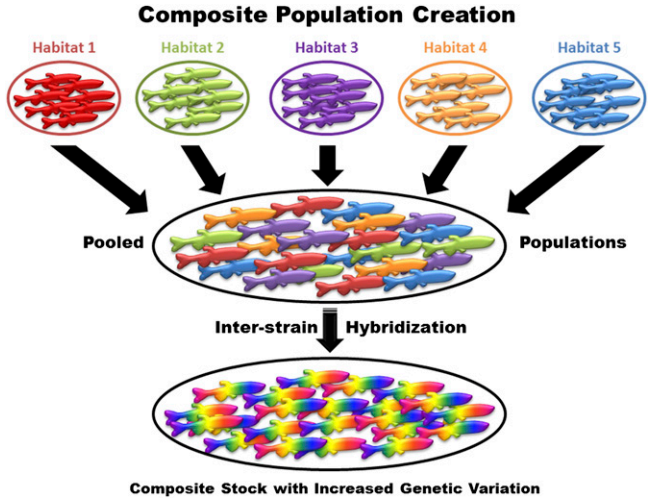

Fig. 5. Composite stocks are created by combining individuals from multiple locally adapted populations. Combined stocks then hybridized in a random intraspecific manner, increasing genetic variability through extensive recombination between divergent populations. In nature, such hybrids are frequently selected against, but in a commercial fish farm lacking selective pressure, hybridized individuals survive and reproduce randomly within the composite population, further increasing genetic variation. Tu, having been created from such a composite population and without significant selective measures, exhibits increased CNV numbers and sizes.

(complete population differentiation), with high V_{st} values indicating regions under increased selective pressure (*SI Text*) (10, 32). Using a stringent threshold cutoff of $V_{st} > 0.6$ (paired average $V_{st} = 0.226$), a total of 678 CNVEs (189 strain-specific) were identified above the V_{st} threshold (Fig. S3*A*). Among laboratory strains, Tu exhibited the highest number of genomic regions with strain-specific V_{st} values greater than the threshold (i.e., CNV V_{st} elevated in one strain versus the other two strains; Tu, 42; WIK, 33; AB, 30). Examples of genomic regions with V_{st} values exclusively elevated in Tu include variants overlapping *rgs4* and *rgs5a* (Fig. S3*B*). These genes have important functions during early hindbrain development and in the adult eye, respectively. Additionally, 84 CNVEs with V_{st} values above the threshold were native fish-specific, perhaps representing increased selective pressures exerted upon the native zebrafish population. Finding fewer strain-specific genomic regions in laboratory strains (Fig. 2), compared with the native population, may be indicative of either inbreeding effects or relaxed selective pressures in laboratory strains compared with natural "native" populations. The likelihood that differences in the native population are the result of relaxed selective pressure in laboratory strains is bolstered by the fact that CNVs among the native fish were significantly smaller than among all of the laboratory strains. This finding is consistent with the idea that larger CNVs, which encompass more sequence and thus are potentially more deleterious, are selectively eliminated from native populations whereas maintained in laboratory strains. Additional analyses would be required to conclusively determine the relative contributions of selective pressure and inbreeding on the observed CNV differences between these zebrafish strains.

The level of variation in native fish was surprisingly lower than that seen in Tu, despite their higher effective population size and presumed ability to migrate and mate randomly. This lower level of variation likely results from the single geographical origin of these fish and the higher selective pressures upon them compared with laboratory strains, as indicated from our V_{st} data. Because AB, WIK, and native fish have similar CNV levels despite their differences in effective population sizes, our findings suggest that newly formed CNVs in native fish genomes may be rapidly eliminated as the result of increased selective pressures.

Implications for Zebrafish Research. The extensive copy number variation observed in our study highlights the need to consider all forms of genetic variation in biological and medical research using zebrafish. Our data indicate not only a high degree of strain substructuring but also an increased level of variation between individual fish within a strain (Table S2). These high levels of variation within zebrafish strains potentially could confound studies intended for translation to human diseases. This possibility is especially important, given we already know of human CNVs that contribute substantially to physiological and phenotypic effects as well as degrees of disease susceptibility and therapy outcomes (12, 33).

It is likely that cellular-level differences arise from some of the zebrafish copy number variants described in this study. For example, it has recently been shown that zebrafish strains exhibit different phenotypic effects when exposed to ethanol (34). Based on the phenotypes examined, individual strains exhibited differences in sensitivity to ethanol with regard to each pathway examined. This finding suggests that ethanol influences each examined phenotypic pathway in a strain-specific manner. Knowledge of CNVs in these pathways may provide candidate genes upon which ethanol acts to result in the various phenotypes.

Summary and Future Directions. This study represents an initial high-resolution CNV map for zebrafish and has identified extensive variation within and among zebrafish strains. (The data from this study are being made publicly available through the Genome Reference Consortium.) This high level of copy number variation among strains has led to a substantial degree of strain substructuring with the potential to cause significant amounts of basal phenotypic differences. This substructuring likely originated from the unique origins of the different zebrafish groups. The decreased number of CNVs and smaller mean CNV size found in the native population, despite its high number of randomly mating individuals, may result from increased selective pressures experienced by natural populations. Our analyses also indicated that intergenic CNVs may have the ability to alter gene expression through both positive and negative interactions.

Integrating copy number variant information into the zebrafish reference genome will enhance future annotation of the reference sequence, especially near sequence gaps and segmental duplications, which often associate with structural variants in the human genome (7). Despite our development and use of this high-resolution zebrafish aCGH platform, CNVs smaller than ~4 kb and other types of genetic variants (i.e., balanced rearrangements and mobile elements) remain undiscovered in zebrafish. Further analyses to uncover these remaining structural genomic variants in the genome may include the use of next-generation sequencing, which would provide nucleotide-level breakpoint information and delineate which mechanisms predominate in CNV formation in the different strains (35, 36). Using complementary technologies to meticulously identify and accurately genotype all forms of genetic variants will ultimately limit the variability in phenotypic outcomes resulting from epistatic effects of background genetic variants. With this knowledge, specific experimental modifications will lead to a more efficient use of zebrafish as an effective model for studies intended for translation to human diseases.

Materials and Methods

Sample Collection and Preparation. Twenty adult zebrafish were collected for the AB (10 from the Dana Farber Cancer Center, Look laboratory; 10 from Purdue University, Freeman laboratory), Tu (10 from Children's Hospital Boston, Zon laboratory; 10 from the University of Utah, Trede laboratory), and WIK (10 from Children's Hospital Boston, Zon laboratory; 10 from the Zebrafish International Resource Center) strains and a native population from Bangladesh (University of Exeter, Tyler laboratory; *SI Text*). All fish were maintained following standard laboratory procedures and euthanized following approved Institutional Animal Care and Use Committee protocols. Euthanized animals were flash-frozen with liquid nitrogen and homogenized for total genomic DNA extraction. Fin clips were subjected to similar freezing and homogenization. Genomic DNA was isolated using a standard phenol-chloroform protocol.

Array Platform Design and Hybridization. All aCGH was performed using a custom-designed Agilent Technologies SurePrint G3 CGH microarray. Uniquely mapping 60-mer oligonucleotide array features were generated using an algorithm that designed probes using the zebrafish Zv8 reference genome. Unique probes were selected to reduce noise generated from highly repetitive regions (i.e., segmental duplications, LINEs, etc.), and 967,331 designed features accompanied 6,685 built-in positive and negative controls, providing an average 1.4-kb probe spacing throughout the zebrafish reference genome. For individual aCGH experiments, one individual from each strain was randomly chosen as a reference sample to compare against all other individuals (test samples) from that strain. Pooled aCGH experiments used equal amounts of DNA from 10 fish from each of the three laboratory strains. Pooled AB DNA was used as the reference against both pooled Tu DNA and pooled WIK DNA, and pooled WIK DNA was used as the reference against pooled Tu DNA. Arrays were hybridized using standard Agilent protocols with one modification: We used 1 μg of heat-denatured DNA (5 min at 95 °C) per labeling reaction in place of 1 μg restriction enzyme-digested DNA. Hybridized arrays were scanned on an Agilent G2505C scanner at 2-μm resolution.

CNV Calling. Array images were extracted using Agilent Feature Extraction software incorporating signal normalization for Cy3 and Cy5 signal intensities. Normalized signal intensity files were imported into and analyzed using Nexus Copy Number software (version 5.1) (BioDiscovery). This program analyzes $\log_2$ ratio output files using a rank segmentation algorithm similar to circular binary segmentation (37). Settings were optimized using a self–self hybridization. Analysis settings can be found in *SI Text*.

Reciprocal Overlap. Detected CNVs were combined to create a nonredundant set of CNVEs. Genetic variation and noise caused by technical differences

between experiments led to a variety of breakpoint coordinates being reported for the same CNV. Thus, we used a 50% reciprocal overlap rule using custom PERL scripts to combine calls sorted by size (smallest to largest), merging those overlapping each other by at least 50% of their respective total lengths.

CNV Validation. Validations of CNV regions were performed using qPCR. Each CNV was validated on the reference sample, one fish presenting the same copy number status as the reference and two fish exhibiting copy number variability. Primers were designed using Primer3 (38) to amplify a 100- to 200-bp fragment within each CNV using sequences from the University of California Santa Cruz (UCSC) Genome Browser (39). A total of 100 CNV regions were randomly chosen for validation (Table S1). Nine control primer pairs were designed and evaluated in ultraconserved elements (40). One control primer pair, (forward) 5′-CCTTTTCCGATGCTTTTTACAC-3′ and (reverse) 5′-GGAAGCCTAGTGCAGTGCTAGT-3′, was selected and amplified in parallel for each sample. qPCR was performed in triplicate using Power SYBR Green PCR Master Mix (Applied Biosystems) in 10-μL reaction volumes on 384-well plates with a 7900HT Real-Time PCR System (Applied Biosystems). The amplification profile consisted of initial activation of AmpliTaq Gold polymerase (Applied Biosystems) at 95 °C (10 min), and then 35 cycles of 95 °C (15 s) and 60 °C (30 s), with dissociation curves generated at PCR completion to confirm the specificity of PCR products. The cycle threshold was determined as the number of cycles needed to cross the threshold value. ΔCt was calculated by subtracting the ultraconserved elements (UCE) Ct value from the CNV Ct value. ΔΔCt was then determined by subtracting the reference sample ΔCt from the test ΔCt. The $\log_2$ ratio as expressed by ΔΔCt for each CNV was then compared with the aCGH $\log_2$ ratio.

CNV Enrichment and Population Genetic Analyses. Randomization tests were performed as described previously (8). Briefly, locations of the 6,080 CNVEs were assigned to the midpoint of a randomly selected probe on the Agilent array, and direct overlap with a zebrafish RefSeq gene was evaluated for 10,000 randomizations of CNVE sizes found in our analysis (*SI Text*). For enrichment analysis, RefSeq gene and assembly gap locations were extracted from the zebrafish reference genome using the "Table" function of the UCSC Genome Browser (39).

Examination of strain-specific genes directly affected by CNV was carried out using Nexus enrichment analysis and gene ontology (GO) analysis (41, 42). GO terms enriched with high degrees of copy number change across the genome were ranked without the need to select regions or place arbitrary thresholds. Potential functional effects were assigned based on this analysis.

V_{st} values were calculated for merged CNVEs using the method described by Redon et al. (10) with modifications. The mean $\log_2$ ratio across all probes falling within a specific CNV region was calculated. The variance of the means for the entire set (V_t), the AB set (V_{AB}), Tu set (V_{Tu}), WIK set (V_{WIK}), and Native set (V_{Native}) was then calculated. Average variance within populations was then calculated (V_s) by taking the mean between populations (i.e., V_{AB} and V_{Tu}). V_{st} values were finally calculated using the standard formula $V_{st} = (V_t - V_s)/V_t$.

ACKNOWLEDGMENTS. We thank S. Stephen and J. Mattick for UCE chromosome locations; Agilent Technologies for zebrafish HD probe sequences; M. Gutierrez-Arcelus, C. Ihm, C. Tran, and I. Smolina for technical assistance; J. Dobrinski for assistance with Figs. 1 and 3; and G. Paull, N. Trede, K. Frazer, T. Look, and C. Tyler for fish. Research was supported by an Institutional National Research Service Award Fellowship (to K.H.B.) from Department of Pathology, Brigham and Women's Hospital and by Grant T32HL007627 [National Institutes of Health (NIH)–National Heart, Lung, and Blood Institute]; Grant 1K99ES018892 (NIH–National Institute of Environmental Health Sciences) (to K.H.B.); Hong Kong General Research Fund (to K.W.C.); and Grants 5R01CA111560 (NIH–National Cancer Institute) and 5P41HG004421 (NIH–National Human Genome Research Institute) (to C.L.).

1. Lieschke GJ, Currie PD (2007) Animal models of human disease: Zebrafish swim into view. *Nat Rev Genet* 8:353–367.
2. Bradley KM, et al. (2007) A major zebrafish polymorphism resource for genetic mapping. *Genome Biol* 8:R55.
3. Coe TS, et al. (2009) Genetic variation in strains of zebrafish (*Danio rerio*) and the implications for ecotoxicology studies. *Ecotoxicology* 18(1):144–150.
4. Conrad DF, et al.; Wellcome Trust Case Control Consortium (2010) Origins and functional impact of copy number variation in the human genome. *Nature* 464:704–712.
5. Egan CM, Sridhar S, Wigler M, Hall IM (2007) Recurrent DNA copy number variation in the laboratory mouse. *Nat Genet* 39:1384–1389.
6. Iafrate AJ, et al. (2004) Detection of large-scale variation in the human genome. *Nat Genet* 36:949–951.
7. Kidd JM, et al. (2008) Mapping and sequencing of structural variation from eight human genomes. *Nature* 453(7191):56–64.
8. Lee AS, et al. (2008) Analysis of copy number variation in the rhesus macaque genome identifies candidate loci for evolutionary and human disease studies. *Hum Mol Genet* 17:1127–1136.
9. Perry GH, et al. (2006) Hotspots for copy number variation in chimpanzees and humans. *Proc Natl Acad Sci USA* 103:8006–8011.
10. Redon R, et al. (2006) Global variation in copy number in the human genome. *Nature* 444:444–454.
11. Emerson JJ, Cardoso-Moreira M, Borevitz JO, Long M (2008) Natural selection shapes genome-wide patterns of copy-number polymorphism in *Drosophila melanogaster*. *Science* 320:1629–1631.
12. Glessner JT, et al. (2009) Autism genome-wide copy number variation reveals ubiquitin and neuronal genes. *Nature* 459:569–573.
13. Bochukova EG, et al. (2010) Large, rare chromosomal deletions associated with severe early-onset obesity. *Nature* 463:666–670.
14. Shlien A, et al. (2008) Excessive genomic DNA copy number variation in the Li-Fraumeni cancer predisposition syndrome. *Proc Natl Acad Sci USA* 105:11264–11269.
15. Jackson AN, McLure CA, Dawkins RL, Keating PJ (2007) Mannose binding lectin (MBL) copy number polymorphism in zebrafish (*D. rerio*) and identification of haplotypes resistant to *L. anguillarum*. *Immunogenetics* 59:861–872.
16. Brownlie A, et al. (2003) Characterization of embryonic globin genes of the zebrafish. *Dev Biol* 255(1):48–61.
17. Rauch G-J, Granato M, Haffter P (1997) A polymorphic zebrafish line for genetic mapping using SSLPs on high-percentage agarose gels. *Tech Tips Online* 2:148–150.
18. Guryev V, et al. (2006) Genetic variation in the zebrafish. *Genome Res* 16:491–497.
19. Sharp AJ, et al. (2005) Segmental duplications and copy-number variation in the human genome. *Am J Hum Genet* 77(1):78–88.
20. Guryev V, et al. (2008) Distribution and functional impact of DNA copy number variation in the rat. *Nat Genet* 40:538–545.
21. Nguyen DQ, Webber C, Ponting CP (2006) Bias of selection on human copy-number variants. *PLoS Genet* 2:e20.
22. Stranger BE, et al. (2007) Population genomics of human gene expression. *Nat Genet* 39:1217–1224.
23. Storey JD (2002) A direct approach to false discovery rates. *J R Stat Soc Series B Stat Methodol* 64:479–498.
24. Zuniga AE, et al. (2004) Mouse limb deformity mutations disrupt a global control region within the large regulatory landscape required for Gremlin expression. *Genes Dev* 18:1553–1564.
25. Engeszer RE, Patterson LB, Rao AA, Parichy DM (2007) Zebrafish in the wild: A review of natural history and new notes from the field. *Zebrafish* 4(1):21–40.
26. Austin JD, Jelks HL, Tate B, Johnson AR, Jordan F (2011) Population genetic structure and conservation genetics of threatened Okaloosa darters (*Etheostoma okaloosae*). *Conserv Genet* 12:981–989.
27. Blouin MS, Phillipsen IC, Monsen KJ (2010) Population structure and conservation genetics of the Oregon spotted frog, *Rana pretiosa*. *Conserv Genet* 11:2179–2194.
28. Tang H, Peng J, Wang P, Risch NJ (2005) Estimation of individual admixture: Analytical and study design considerations. *Genet Epidemiol* 28:289–301.
29. Mullins MC, Hammerschmidt M, Haffter P, Nüsslein-Volhard C (1994) Large-scale mutagenesis in the zebrafish: In search of genes controlling development in a vertebrate. *Curr Biol* 4(3):189–202.
30. Verhoeven KJF, Macel M, Wolfe LM, Biere A (2011) Population admixture, biological invasions and the balance between local adaptation and inbreeding depression. *Proc Biol Sci* 278(1702):2–8.
31. Freeman JL, et al. (2007) Definition of the zebrafish genome using flow cytometry and cytogenetic mapping. *BMC Genomics* 8:195.
32. Lawrence C, Ebersole JP, Kesseli RV (2008) Rapid growth and out-crossing promote female development in zebrafish (*Danio rerio*). *Environ Biol Fishes* 81:239–246.
33. McCarroll SA, Altshuler DM (2007) Copy-number variation and association studies of human disease. *Nat Genet* 39(Suppl 7):S37–S42.
34. Loucks E, Carvan MJ, III (2004) Strain-dependent effects of developmental ethanol exposure in zebrafish. *Neurotoxicol Teratol* 26:745–755.
35. Mills RE, et al.; 1000 Genomes Project (2011) Mapping copy number variation by population-scale genome sequencing. *Nature* 470(7332):59–65.
36. Alkan C, Sajjadian S, Eichler EE (2011) Limitations of next-generation genome sequence assembly. *Nat Methods* 8(1):61–65.
37. Olshen AB, Venkatraman ES, Lucito R, Wigler M (2004) Circular binary segmentation for the analysis of array-based DNA copy number data. *Biostatistics* 5:557–572.
38. Rozen S, Skaletsky H (2000) Primer3 on the WWW for general users and for biologist programmers. *Bioinformatics Methods and Protocols: Methods in Molecular Biology*, ed Krawetz SMS (Humana, Totowa, NJ), pp 365–386.
39. Karolchik D, et al. (2004) The UCSC Table Browser data retrieval tool. *Nucleic Acids Res* 32(Database issue):D493–D496.
40. Stephen S, Pheasant M, Makunin IV, Mattick JS (2008) Large-scale appearance of ultraconserved elements in tetrapod genomes and slowdown of the molecular clock. *Mol Biol Evol* 25:402–408.
41. Mootha VK, et al. (2003) PGC-1α-responsive genes involved in oxidative phosphorylation are coordinately downregulated in human diabetes. *Nat Genet* 34:267–273.
42. Subramanian A, et al. (2005) Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. *Proc Natl Acad Sci USA* 102:15545–15550.